# Московский авиационный институт (национальный исследовательский университет)

# Факультет информационных технологий и прикладной математики

# Кафедра вычислительной математики и программирования

# Лабораторная работа №0 по курсу «Искуственный интеллект»

# Студент: К.В. Лукашкин Группа: М8О-308Б

# Москва, 2019

## Получение и обработка данных

## Задача.

Требуется сформировать/получить два набора данных соответствующие следующим критериям:

1. Один из датасетов должен представлять собой корпус документов. Язык, источник и тематика произвольна
2. Второй датасет должен содержать категориальные, количественные признаки. Для данного датасета определить предсказываемые признаки (для задачи регрессии и классфикации). Если такого признака нет, спроектировать

Данные датасеты будут в дальнейшем использованы в оставшихся лабораторных работах.

По каждому датасету построить распределения признаков (в случае корпуса документов – построить распределение слов) и объяснить имеющуюся картину. Вычислить статистические характеристики признаков. Обнаружить и решить возможные проблемы с данными. Если решить данную проблему невозможно, объяснить почему.

## Датасеты.

Программы реализованны на Python3 с использованием библиотеки pandas

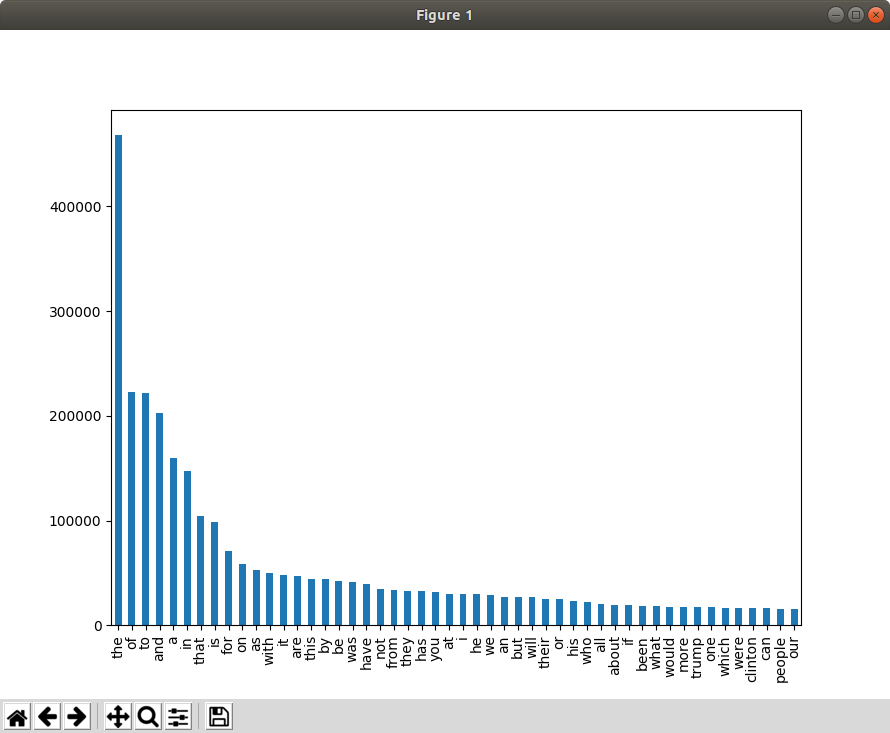
### 1. Корпус документов

Getting Real about Fake News

<https://www.kaggle.com/mrisdal/fake-news>

13000 новостных статей которые могут быть отнесены к категории заведомо ложные или предвзятые.

Код находится в файле lab0\_ai.py



Чаще всего встречаются наиболее употребимые слова – предлоги, артикли, союзы и местоимения. Это свойственно любому ествественному языку.

Стоит отметить наличие среди 50 наиболее часто исползуемых слов – trump и clinton, что соотвествует специфике выборки.

### 2. Датасет с категориальными, количественными признаками

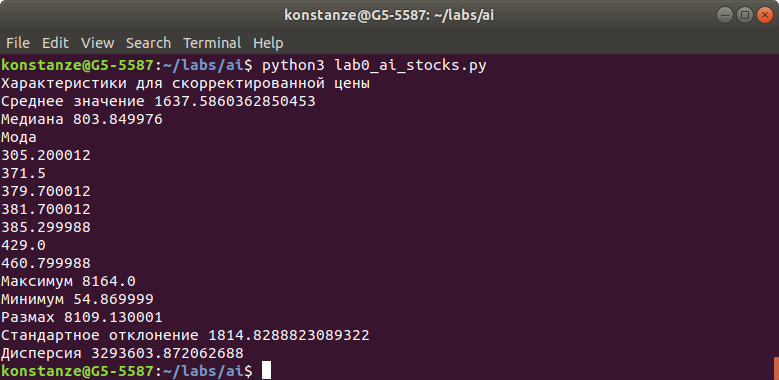
История акций NASDAQ Composite (^IXIC) с самого начал их существования 2/5/1971.

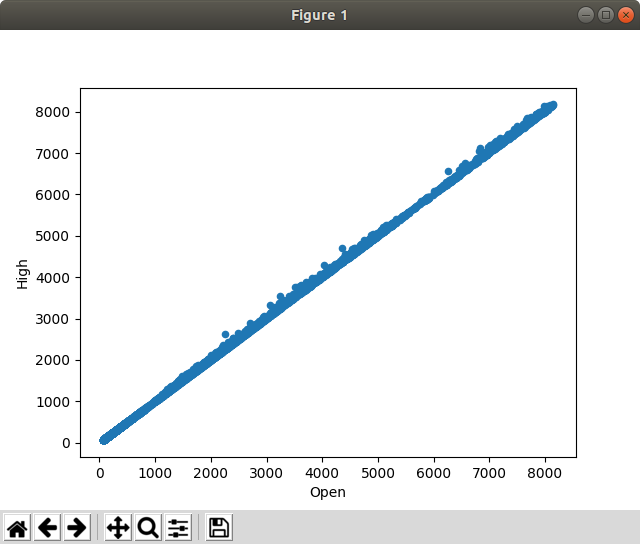
[https://finance.yahoo.com/quote/%5EIXIC/history](https://finance.yahoo.com/quote/^IXIC/history)

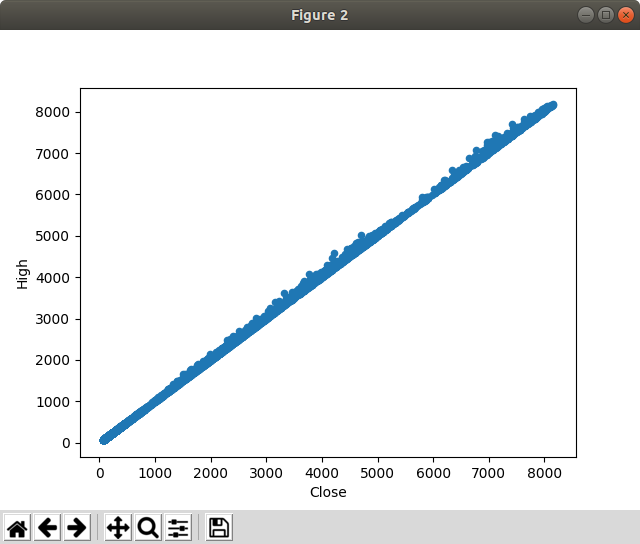
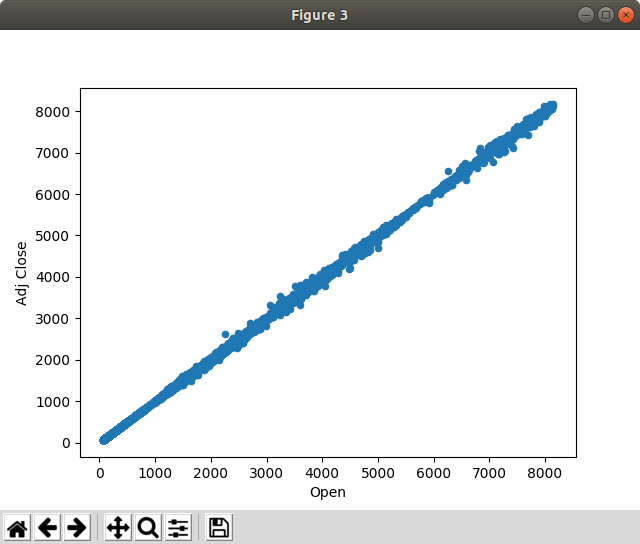
В таблице представлены дата торгового дня, цена при открытии, максимальная, минимальная цены во время торгов, цена при закрытии, скорректированная цена закрытия и объемы акций.

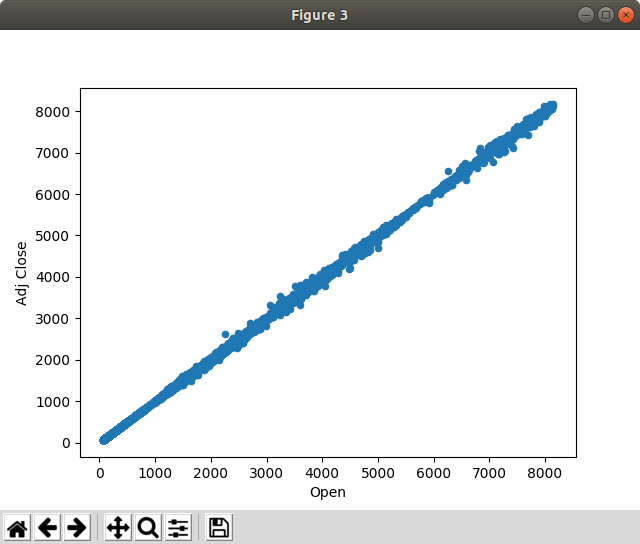
Основные признаки которые требуется предсказать: максимальная/минимальная цена за день.

Код: lab0\_ai\_stocks.py









Распределение признаков:

1. Цена на момент открытия - максимальная цена

Максимальная цена в подавляющем большинстве случаев оказывается осязаемо больше цены открытия – это связано с тем что цена акций совершает довольно большие колебания в течение дня.

1. Цена закрытия - максимальная цена за торговый день.

Цена на момент закрытия редко имеет сильное отличие от максимальной цены за день. Это характерно почти для любой компании, конечно есть периоды когда компания совершает какую-то ошибку на мировой арене, но такие периоды как правило короткие.

1. Цена открытия — скорректировання цена закрытия

Цена закрытия в многих случаях больше цены открытия, это связано с тем, что в такой большой выборке компания вела дела довольно успешно. Однако чётко видно что иногда это не выполняется, это связано с довольно продолжительными кризисами.

Весь код находиться на <https://github.com/memosiki/mai_ai>

## Вывод.

Выполнив лабораторную работу, я ознакомился с библиотекой pandas. Изучил способы нахождения информации и датасетов. Освоил несложную характеризацию и обработку данных.